Adaptive Region-Based Multimodal Image Fusion Using ICA Bases

Nedeljko Cvejic, John Lewis, David Bull, Nishan Canagarajah
Department of Electrical and Electronic Engineering
University of Bristol
Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, United Kingdom
{n.cvejic,john.lewis,david.bull,nishan.canagarajah}@bristol.ac.uk

Abstract - ***In this paper, we present a novel multimodal image fusion algorithm in ICA domain. It uses segmentation to determine the most important regions in the input images and consequently fuses the ICA coefficients from given regions using the Piella fusion metric to maximise the quality of the fused image. The proposed method exhibits significantly higher performance than the basic ICA algorithm and improvement over other state-of-the-art algorithms.***

Keywords: image fusion, region-based image fusion, image fusion metrics, segmentation, Independent Component Analysis.

1 Introduction

Fusion of visible and infrared (IR) images and video sources, is becoming increasingly important for surveillance purposes. A relatively lower level of interest in infrared imagery, compared to visible imagery, has been due to high cost of thermal sensors, lower image resolution, higher image noise and lack of widely available data sets. However, these drawbacks are becoming less relevant as infrared imaging advances, making the technology important for applications such as video surveillance, navigation, face recognition and object tracking. Night vision cameras have also become available that produce images in multiple spectral bands, e.g. thermal and visible. These different bands provide complementary information since they represent different characteristics of a scene or object.

A fused image constructed by combination of the visible and infrared inputs, or some of their features, enables improved detection and unambiguous localisation of a target (represented in the thermal image) with respect to its background (represented in the visible image) [1]. A human operator using a suitably fused representation of visible and IR images may therefore be able to construct a more complete and accurate mental representation of the perceived scene, resulting in a larger degree of situation awareness [2].

The image fusion process can be performed at different levels of information representation: signal, pixel, feature and symbolic level. One of the feature-level fusion methods is the region-based imaged fusion. Images to be fused are initially segmented into a set of distinctive regions. Various properties of the regions obtained by segmentation can be used to determine which features from which images are to be included in the fused image. This has advantages over pixel-based methods as more intelligent semantic fusion rules can be considered based on actual features in the image, rather than on single or arbitrary groups of pixels.

Nikolov et al [3] proposed a classification of image fusion algorithms into spatial domain and transform domain techniques. Instead of using a standard bases system, such as the DFT, the mother wavelet or cosine bases of the DCT, one can train a set of bases that are suitable for a specific type of images. A training set of image patches, which are acquired randomly from images of similar content, can be used to train a set of statistically independent bases. This is known as Independent Component Analysis (ICA) [4]. Recently, several algorithms have been proposed [5, 6], in which ICA and bases are used for transform domain image fusion. In this paper, we refine the approach by a novel multimodal image fusion algorithm in ICA domain. It uses segmentation to determine the most important regions in the input images and consequently fuses the ICA coefficients from given regions using fusion metrics to maximise the quality of the fused image.

2 Background Review

In order to obtain a set of statistically independent bases for image fusion in the ICA domain, training is performed with a predefined set of images. Training images are selected in such a way that the content and statistical properties are similar for the training images and the images to be fused. An input image $i(x, y)$ is randomly windowed using a rectangular window w of size $N \times N$. The result of windowing is an "image patch" which is defined as [5]:

$$p(m, n) = w \cdot i(m_0 - N/2 + m, n_0 - N/2 + n) \quad (1)$$

where m and n take integer values from the interval $[0, N-1]$. Each image patch $p(m, n)$ can be represented by a linear combination of a set of M basis patches $b_i(m, n)$:

$$p(m, n) = \sum_{i=1}^{M} v_i b_i(m, n) \quad (2)$$

where $v_1, v_2, ..., v_M$ stand for the projections of the original image patch on the basis patch, i.e. $v_i = \langle p(m,n), b_i(m,n) \rangle$. A 2D representation of the image patches can be simplified to a 1D representation, using lexicographic ordering. This implies that an image patch $p(m,n)$ is reshaped into a vector $\underline{p}$, mapping all the elements from the image patch matrix to the vector in a row-wise fashion. Decomposition of image patches into a linear combination of basis patches can the be expressed as follows:

$$\underline{p}(t) = \sum_{i=1}^{M} v_i(t)\underline{b}_i = [\underline{b}_1 \underline{b}_2 ... \underline{b}_M] \cdot \begin{bmatrix} v_1(t) \\ v_2(t) \\ ... \\ v_M(t) \end{bmatrix} \quad (3)$$

where t represents the image patch index. If we denote $B = [b_1 b_2 ... b_M]$ and $v(t) = [v_1 v_2 ... v_M]^T$, then equation (3) reduces to:

$$\underline{p}(t) = B\underline{v}(t) \quad (4)$$

$$\underline{v}(t) = B^{-1}\underline{p}(t) = A\underline{p}(t) \quad (5)$$

Thus, $B = [b_1 b_2 ... b_M]^T$ represents an unknown mixing matrix (analysis kernel) and $A = [a_1 a_2 ... a_M]^T$ the unmixing matrix (synthesis kernel). This transform projects the observed signal $\underline{p}(t)$ on a set of basis vectors. The aim is to estimate a finite set of $K < N^2$ basis vectors that will be capable of capturing most of the input image properties and structure.

In the first stage of basis estimation the Principal Component Analysis (PCA) is used for dimensionality reduction. This is obtained by eigenvalue decomposition of the data correlation matrix $C = E\{\underline{p}\,\underline{p}^T\}$. The eigenvalues of the correlation matrix illustrate the significance of their corresponding basis vector. If V is the obtained $K \times N$ PCA matrix, the input image patches are transformed by:

$$\underline{z}(t) = V\underline{p}(t) \quad (6)$$

After the PCA preprocessing step we select the statistically independent basis vectors using the optimisation of the negentropy. The following rule defines a FastICA approach that optimises negentropy, as proposed in [4]:

$$\underline{a}_i^+ \leftarrow \varepsilon\{\underline{a}_i \phi(\underline{a}_i^T \underline{z})\} - \varepsilon\{\phi'(\underline{a}_i^T \underline{z})\}\underline{a}_i \qquad 1 \le i \le K \quad (7)$$

$$A \leftarrow A(A^T A)^{-0.5} \quad (8)$$

where $\phi(x) = -\partial G(x)/\partial x$ defines the statistical properties $G(x) = \log p(x)$ of the signals in the transform domain [4]. In our implementation we used:

$$G(x) = \alpha\sqrt{\eta + x} + \beta \quad (9)$$

where α and β are constants and η is a small constant to tackle numerical instability, in the case that $x \to 0$ [4].

After the input image patches $\underline{p}(t)$ are transformed to their ICA domain representations $\underline{v}_k(t)$, we can perform image fusion in the ICA domain in the same manner as it is performed in e.g. the wavelet domain. The equivalent vectors $\underline{v}_k(t)$ from each image are combined in the ICA domain to obtain a new image $\underline{v}_f(t)$. The method that combines the coefficients in the ICA domain is called the "fusion rule". After the composite image $v_f(t)$ is constructed in the ICA domain, we can move back to the spatial domain, using the synthesis kernel A, and synthesise the image $i_f(x,y)$. Several features can be employed in the estimation of the contribution of each input image to the fused output image. In [5], the authors proposed the mean absolute value of each $N \times N$patch in the transform domain, as an activity indicator:

$$E_k(t) = \|\underline{v}_k(t)\|, \qquad k = 1, ..., T \quad (10)$$

where T denotes the number of input images. As the ICA bases tend to focus on the edge information, large values for $E_k(t)$ correspond to increased activity in the patch, e.g. the existence of edges. Based on this observation, the authors in [5] divide the transform domain patches in two groups. The first group consists of the regions that contain details ($E_k(t)$ larger then a threshold) and the second group contains the region with background information ($E_k(t)$ smaller then a threshold). The threshold that determines whether a region is "active" or "non-active" is set heuristically. As a result, the segmentation map $s_k(t)$ is created for each input image [5]:

$$s_k(t) = \begin{cases} 1 & \text{if } E_k(t) > \frac{2}{T}\sum_{k=1}^{T} E_k(t) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

The segmentation maps of input images are combined to form a single segmentation map, using the logical OR operator [5]:

$$s(t) = OR\{s_1(t), s_2(t), ..., s_T(t)\} \quad (12)$$

After the input images are segmented into active and non-active regions, two different fusion rules are used for fusion of each group of regions [5]. Namely, active regions are fused using the "max-abs" rule, while non-active regions are fused using the "mean" rule. The "max-abs" rule fuses two input coefficients/vectors by selecting the one with higher absolute value. In the "mean" fusion rule the fused coefficient/vector is equal to the mean value of the two input coefficients/vectors.

3 Proposed Method

In this paper we focus on the fusion of infra-red (IR) and visible images, although methods can be generalized to other modalities. Because the threshold that determines the "activity" of a region is set heuristically, the regions obtained by thresholding of the ICA coefficients do not correspond always to objects in the images to be fused. Our experiments showed that important objects in the IR input images (e.g. a person or a smaller object) are often masked by textured high-energy background in the visual image. In this case the important objects from the IR image become blurred or, in extreme cases, completely masked. Therefore, we perform segmentation in the spatial domain and then

fuse patches from separate regions separately. This differs from the methods in [5, 6] where the fusion was performed on a more general, pixel level.

3.1 The segmentation algorithm

The quality of the segmentation algorithm is of vital importance to the fusion process. An adapted version of the combined morphological−spectral unsupervised image segmentation algorithm is used, which is described in [7], enabling it to handle multi-modal images. The algorithm works in two stages. The first stage produces an initial segmentation by using both textured and non-textured regions. The detail coefficients of the DT-CWT are used to process texture. The gradient function is applied to all levels and orientations of the DT-CWT coefficients and up-sampled to be combined with the gradient of the intensity information to give a perceptual gradient. The larger gradients indicate possible edge locations. The watershed transform of the perceptual gradient gives an initial segmentation. The second stage uses these primitive regions to produce a graph representation of the image which is processed using a spectral clustering technique.

The method can use either intensity information or textural information or both to obtain the segmentation map. This flexibility is useful for multi-modal fusion where some a priori information of the sensor types is known. For example, IR images tend to lack textural information with most features having a similar intensity value throughout the region. Therefore, we used an intensity only segmentation map, as it gives better results than a texture based segmentation.

The segmentation can be performed either separately or jointly. For separate segmentation, each of the input images generates an independent segmentation map for each image.

$$S_1 = \sigma(i_1, D_1), \ldots, S_N = \sigma(i_N, D_N) \qquad (13)$$

where D_n represent detail coefficients of the DT-CWT used in segmentation. Alternatively, information from all images could be used to produce a joint segmentation map.

$$S_{joint} = \sigma(i_1 \cdots i_N, D_1 \cdots D_N) \qquad (14)$$

In general, jointly segmented images work better for fusion [8]. This is because the segmentation map will contain a minimum number of regions to represent all the features in the scene most efficiently. A problem can occur for separately segmented images, where different images have different features or features which appear as slightly different sizes in different modalities. Where regions partially overlap, if the overlapped region is incorrectly dealt with, artefacts will be introduced and the extra regions created to deal with the overlap will increase the time taken to fuse the images.

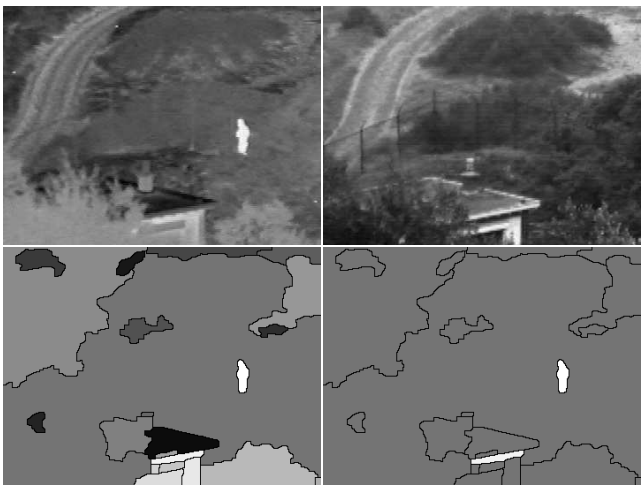

Figure 1: Segmentation and region selection prior to fusion. (a) IR input image, (b) Visible input image, (c) Regions obtained by joint segmentation of the input images and (d) The image mask: white from IR, gray from visible.

3.1.1 Calculation of priority and fusion rules

After the images are jointly segmented it is essential to determine the importance of regions in each of the input images. We have decided to use the normalized Shannon entropy of a region as the priority. Thus, the priority $P(r_{t_n})$ is given as:

$$P(r_{t_n}) = \frac{1}{|r_{t_n}|} \sum_{\forall\theta,\forall l,(x,y)\in r_{t_n}} d^2_{n(\theta,l)}(x,y) \log d^2_{n(\theta,l)}(x,y) \qquad (15)$$

with the convention $0\log(0) = 0$, where $|r_{t_n}|$ is the size of the region r_{t_n} in input image n and $d_{n(\theta,l)}(x,y) \in D_{n(\theta,l)}$ detail coefficients of the DT-CWT used in segmentation. Finally, a mask M is generated that determines which image each region should come from in the fused image. An example of the IR input image, visual input image, performed joint segmentation and the image fusion mask is given in Figure 1.

3.2 Weighted Image Reconstruction using Fusion Metrics

After image fusion masks have been determined, weighted region-based image fusion is performed, using the Piella fusion metric [9]. The main aim is to transfer all the important regions from the IR image to the fused image, while retaining the background details from the visible image, thus increasing the perceptual quality of the fused image. This approach can be extended to different modalities and more than two modalities. We propose a novel method for reconstruction of the fused image, using statistical properties of the both input images. In the standard ICA method [5], reconstruction of the fused image is performed on the patch-per-patch base:

$$F(i) = U(i) + V_1(i) + V_2(i) \qquad (16)$$

where $F(i)$ represents the i-th patch of the fused image $i_f(x,y)$, $U(i)$ is the i-th patch obtained by inverse

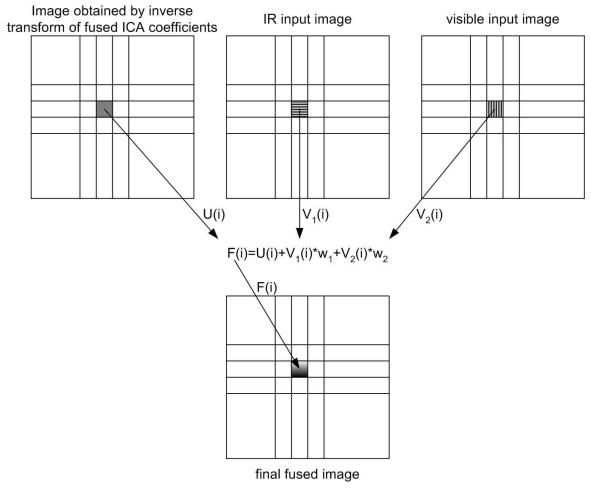

Figure 2: Overview of the proposed weighted ICA fusion method

transform of the fused ICA coefficients and $V_1(i)$ and $V_2(i)$ are the mean values of the corresponding patches from $i_1(x, y)$ and $i_2(x, y)$, respectively, as shown in Figure 2. We propose a new approach for reconstruction of the fused image:

$$F(i) = U(i) + V_1(i) \cdot w_1 + V_2(i) \cdot w_2 \qquad (17)$$

where weights $w_1 \in [0, 1]$ and $w_2 (= 1 - w_1) \in [0, 1]$ are used to balance the contributions from both visual and IR images in the synthesis of the fused image. Weighting coefficients are set to a predefined value (e.g. $w_1 = 1$ and $w_2 = 0$) and then gradually increased or decreased. The Piella image fusion metric [9] is calculated at each step using 8×8 windows from input images and the temporary fused image at the given region. When the maximum value of the Piella metric is reached, the process stops and reconstruction of the fused image is performed with the calculated weights. In that sense, the weighting coefficients are calculated for each 8×8 window so that the quality of the fused image is maximized. The adaptive fused image reconstruction adds only $1 - 2\%$ of computational overhead to the standard, non-adaptive algorithm. Figure 3 depicts the impact of the weighting factors on the visual quality of a fused image in terms of the Piella metric. It is clear that the introduced adaptivity in the reconstruction of the fused image significantly improves the performance of the proposed algorithm.

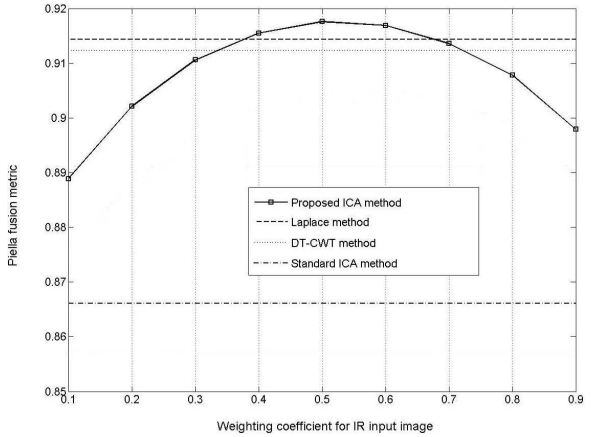

Figure 3: Proposed weighting method in the reconstruction of the fused image, using the Piella image fusion metric

4 Experimental Results

The proposed image fusion method was tested in the multimodal scenario with two input images: infrared and visible. In order to make a comparison between the proposed method and the standard ICA method, the images were fused using the approach described in [5]. We compared these results with a simple averaging method, the Laplace transform (LT) and the dual-tree complex wavelet transform (DT-CWT)[1]. In the multiresolution methods (LT, DT-CWT) a 5-level decomposition is used and fusion is performed by selecting the coefficient with a maximum absolute value, except for the case of the lowest resolution subband where the mean value is used. Before performing image fusion,

Figure 4: Fusion results. Top: input visible image (left), input IR image (right), Middle: fused image using standard ICA fusion (left), fused image using DT-CWT (right), Bottom: fused image using LT (left), fused image using proposed ICA method (right)

the ICA bases were trained using a set of images with content comparable to the test set. The number of rectangular patches ($N = 8$) used for training was 10000, randomly selected from the training set. The lexicographic ordering was applied to the image patches and then PCA performed. Following this, the 32 most important bases ($K = 32$) were selected, according to the eigenvalues corresponding to these bases. After that, the ICA update rule in (7) was iterated until convergence. ICA coefficients are combined using the princi-

Figure 5: Fusion results. Top: input visible image (left), input IR image (right), Middle: fused image using standard ICA fusion (left), fused image using DT-CWT (right), Bottom: fused image using LT (left), fused image using proposed ICA method (right)

ple described in Section 2, while reconstruction of the fused image was performed using optimisation based on the Piella fusion performance metric [9].

Example input images and fused outputs are given in Figure 4 and Figure 5. Visual (subjective) comparison between methods indicates that our method is far superior to the basic ICA method, but also that the proposed weighted ICA method performs slightly better than the LT and DT-CWT methods: for example, in Figure 4 it is clear that the fence detail from the visual image is far better transferred into the fused image in the proposed method than in the standard ICA method. In addition, the details of the two trees in the visual image are visually more pleasing in the proposed method than in the DT-CWT approach, although the person is brighter in the DT-CWT fused image. In Figure 5 it is clear that the DT-CWT method obtains slightly better subjective quality, although it scores lower in the fusion metrics. However, it should be noted that it is the only test image with dimensions 640×480, whereas other test images were 360×270 (the same number of training patches was used - 10000). If the number of training patches was increased in accordance with the size of the images, the subjective quality would likely improve.

The data presented in Table 1 and Table 2 confirms the visual (subjective) fusion assessment, using both the Petrovic [10] and the Piella metric [9]. The proposed method exhibits significantly higher performance than the basic ICA algorithm and improvement over other state-of-the-art algorithms.

Finally, it is important to note that, although

Table 1: Performance of the image fusion methods measured by standard fusion metrics.

Image number	UN Camp 1812		Octec 04	
Metric/Method	Piella	Petrovic	Piella	Petrovic
Average	0.8570	0.3478	0.8902	0.4809
Laplace	0.9144	0.5011	0.9505	0.7257
DT-CWT	0.9123	0.4622	0.9510	0.7247
ICA	0.8661	0.3782	0.8510	0.5225
proposed ICA	**0.9292**	**0.5400**	**0.9515**	**0.7821**

Table 2: Performance of the image fusion methods measured by standard fusion metrics.

Image number	Dune 7404		Trees 4917	
Metric/Method	Piella	Petrovic	Piella	Petrovic
Average	0.9622	0.5139	0.9096	0.4448
Laplace	0.9705	0.5998	0.9266	0.5541
DT-CWT	0.9713	0.6003	0.9270	0.5540
ICA	0.9633	0.5330	0.8409	0.4444
proposed ICA	**0.9741**	**0.6515**	**0.9359**	**0.5829**

training-based algorithms can offer improved performance in specific scenarios where contextual information is available, untrained algorithms such as the DT-CWT still offer a powerful alternative for many applications.

5 Conclusions

We have presented a new multimodal image fusion algorithm in ICA domain. It uses weighting of the ICA bases during reconstruction of the fused image and optimizes its quality using the Piella fusion performance metric. Experimental results confirm that the proposed method exhibits significantly better fusion than basic ICA method, as it obtains higher scores using both Piella and Petrovic metrics.

References

[1] A. Toet, J. K. IJspeert, A. M. Waxman and M. Aguilar, Perceptual evaluation of different image fusion schemes, *Displays*, 24:25–37, 2003.

[2] A. Toet and E. M. Franken, Fusion of visible and thermal imagery improves situational awareness, *Displays*, 18:85–95, 1997.

[3] S. Nikolov, P. Hill, D. Bull, and N. Canagarajah, *"Wavelets for image fusion", in Wavelets in Signal and Image Analysis*, Kluwer, Dordrecht, The Netherlands, 2001.

[4] A. Hyvrinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, London, United Kingdom, 2001.

[5] N. Mitianoudis, T. Stathaki, Pixel-based and Region-based Image Fusion schemes using ICA bases, *Information Fusion*, to appear, 2006.

[6] N. Cvejic, D. Bull and N. Canagarajah, A novel ICA domain multimodal image fusion algorithm, *Proc. SPIE Defense and Security Symposium*, Orlando, FL, to appear, 2006.

[7] R. O'Callaghan and D. Bull, Combined morphologicalspectral unsupervised image segmentation, *IEEE Transactions on Image Processing*, 14:49-62, 2005.

[8] J. Lewis and R. O'Callaghan and S. Nikolov and D. Bull and N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, *Information Fusion*, to appear, 2006.

[9] G. Piella and H. Heijmans, A new quality metric for image fusion, *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, 173–176, 2003.

[10] C. Xydeas and V. Petrovic, Objective pixel-level image fusion performance measure, *Proc. SPIE*, Orlando, FL, 88–89, 2000.